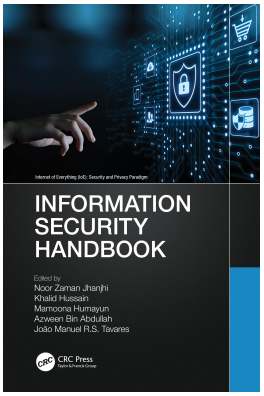

Information Security Handbook

Noor Zaman Jhanjhi, Khalid Hussain, Azween Bin Abdullah, Mamoona Humayun, João Manuel R.S. Tavares

Security in Big Data

Publication details
https://test.routledgehandbooks.com/doi/10.1201/9780367808228-4
Mehwish Malik, Hina Umbrin, Nuzhat Akram, Khalid Hussain Usmani, NZ Jhanjhi
Published online on: 18 Feb 2022

4 Security in Big Data

Mehwish Malik, Hina Umbrin, Nuzhat Akram, Khalid Hussain Usmani, and NZ Jhanjhi

CONTENTS

4.1 Big Data 56
4.1.1 Volume – The Size of Data 58
4.1.2 Variety – Different Forms of Data 59
4.1.3 Velocity – Speed of Data Generation 60
4.1.4 Value – Data's Worth 61
4.1.5 Veracity – Data Uncertainty 61
4.1.6 Variability – Data Inconsistency 62
4.1.7 Visualization – Data Representation 62
4.1.8 Volatility – How Long to Store Data 62
4.1.9 Validity – Data Use 63
4.2 Data Sources of Big Data 63
4.3 Architecture 64
4.3.1 Data Source 64
4.3.2 Ingestion 64
4.3.3 Storage Layer 64
4.3.4 Staging 64
4.3.5 Data Pipeline 65
4.3.6 Data and Workflow Management 65
4.3.7 Data Access 65
4.4 Big Data Challenges 65
4.5 Big Data Analytics Challenges in Big Data 66
4.6 Technical Challenges in Big Data 68
4.7 Characteristics-Oriented Challenges of Big Data 69
4.7.1 Data Volume 69
4.7.2 Data Velocity 69
4.7.3 Data Variety 70
4.7.4 Data Value 70
4.8 Privacy Challenges 70
4.9 Security Challenges 71
4.10 Current Security Challenges in Big Data 71
4.10.1 Big Data Security – A Definition 71
4.10.2 Case Studies of Security Breaches Depicting Their Impact on Organizations 72

DOI: 10.1201/9780367808228-4

4.11 Major Security Issues of Big Data 73
4.11.1 Distributed Frameworks Security 73
4.11.2 Nonrelational Data Stores Protection 74
4.11.3 Storage Security 74
4.11.4 Monitoring Real-Time Security 75
4.11.5 Privacy-Preserving Data Analytics and Mining 75
4.11.6 Granular Audit 75
4.11.7 End-Point Security 76
4.11.8 Data-Centric Security Based on Cryptography 76
4.12 Solutions to Security Challenges 76
4.12.1 Complete Data Supervision of Social Networks 77
4.12.2 Improvement in Legal Mechanism 77
4.12.3 Improvement to People Awareness of Data Quality 77
4.12.4 Put Security First 77
4.13 Conclusion 77
References 78

4.1 BIG DATA

To begin with, it is a common belief that big data describes huge data sets that need novelties in analytical methods to fully utilize them and create new and innovative kinds of value. The vastness of big data is not because of absolute magnitude or size; instead, it's about the appropriate scale of studies and analysis. Researchers have defined it in various ways since it's a ubiquitous term exploited in various parts of academia and industries. Sagiroglu [1] associated big data to volumes of data sets, saying big data is an expression for large-scale data sets having huge, more diverse, as well as complex, structures with the complications of analysing, storing, and visualizing data for additional processing and results. Another definition proposed by Van Dijck [2] stated big data as social action transformation to online measured data, which consequently allow predictive analysis and real-time tracking. They encourage aspirations to build more reliable and accurate predictions to resolve complex and intricate problems, ranging from changes in climate to terrorist activities Kitchin [3]. Furthermore, big data embody administrative challenge regarding extensive information accumulation by corporate ventures, as well as state agencies. Bekker [4] described big data as the data characterized by informational features, such as statistical correctness and the nature of event logs, etc., and that urges such technical necessities as parallel processing, distributed storage, and uncomplicated solution scalability. The author further described each feature comprehensively, arguing that traditional data was suspectable to change at any time; for instance, bank accounts, product counts at a warehouse, etc., whereas big data depicts a log of each record denoting certain events, for example, web page view, purchase activity in a store, sensor value w.r.t time period, social media comments, etc. This nature of big data allows data of events to not change. Though big data is considered statistically correct, it may contain errors or omissions, which is why it is not considered a good choice for tasks requiring absolute accuracy. Another property of big data highlighted

by the author is its technical requirements due to its volume; it requires parallel processing, and high-storage capacity needs a special storage approach.

Andrea [5] identified core themes linked to big data, which are information, method, technology, and its impact. They have quoted various explanations of the term big data and checked if they possess the aforementioned themes or not. Information can be termed as one of the fundamental reasons behind the existence of big data, as its generation and availability serve as fuel to big data. Information in big data gave rise to a new term called datafication, which aims to organize a digital form of analog signals to produce insights that couldn't have been inferred otherwise. Technology is frequently associated with big data, which allows its exploitation such as Hadoop [6,7]. The capability to efficiently store an extensive amount of data on relatively smaller machines is a fundamental element of technology application on big data. Hence, it's justified to say that technology is the core equipment to work with big data. Extensive analysis of quantitative data and a necessity to grab the value from an individual's behavior need processing techniques and methods; thus, methods transform big data into an asset. The exploitation of big data analytics not only allows to manage data efficiently and properly but also help incorporation of such data for decision-making processes. Awareness of these methods, along with technologies, their strength and weaknesses, and cultural tendencies, spread to facilitate informed and intelligent decision making, as required by big data. The degree of impact it is imposing on our society is often portrayed through success stories and anecdotes of technology and methods of implementation. These stories combined, with novel principles and procedural developments, lead to a valuable contribution toward knowledge creation on a subject. If the pervasive quality of information availability and productions results in tons of applications spanning several scientific fields, then it also has an adverse impact on society as well. Major concerns arising due to the evolution of big data are: privacy issues, issues regarding information accessibility, etc.; henceforth, impact is an integral theme of big data. Authors [5] then grouped the definitions into a few groups, the first being the group focusing entirely on enlisting characteristics of big data. In this group, Laney [8] presented a framework that expresses the increase in volume, variety, and velocity of data in all three dimensions; this framework was later named as three V's of big data. This model was later extended to value [9], veracity [10], variability [11], viscosity [12], virality [12], and validity [13,14], described in detail below. Another group defined big data in terms of technical requirements behind processing huge amounts of data. The rest of the definitions associated with big data to crossing certain thresholds, such as when data surpass the processing capacity of traditional database systems, then it can be termed as big data. A conclusive definition was proposed [5] in light of previous definitions; it states that big data is a representation of informational resources characterized by high velocity, volume, and variation to necessitate specific analytical processes and technology to transform it into value. To cap it all, big data is simply a transformation where data is processed into information, which is processed into knowledge, which turns into wisdom and value, as demonstrated in Figure 4.1 (Big Data Transformation).

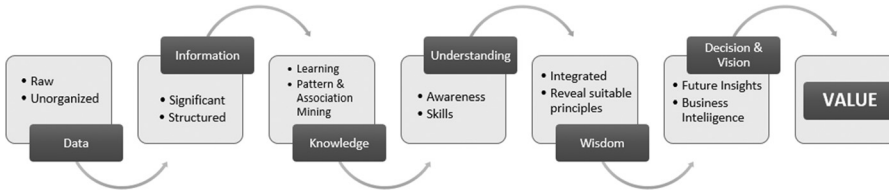

FIGURE 4.1 Big Data Transformation.

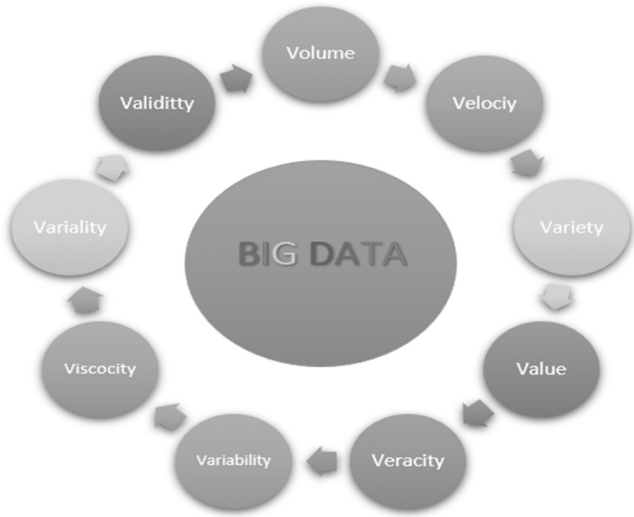

FIGURE 4.2 9 V's of big data.

To sum it all, big data shapes a phenomenon based on technology, culture, and scholar knowledge that stands on the interaction of technology, mythology, and analysis that provokes wide dystopian and utopian rhetoric Boyd [15].

Big data characteristics were initially described by Laney [8] as part of a data framework, not directly associated to big data but was made part of the big data model later and was then extended by several authors. These characteristics are explained below and also illustrated in Figure 4.2.

4.1.1 Volume – The Size of Data

The sheer volume is principle trademark that makes the data 'big.' Volume is surely the base, if we consider big data as a pyramid. It states the enormous amount of data being produced every second from cell phones, social media, M-2-M sensors, credit cards, videos, photographs and so forth. The present quantity of data can be pretty staggering. Consider a few examples:

- In an annual report prepared by Thomas Reuter in 2010, he shared his estimation of world data to be more than 800 exabyte and sill growing further. In 2012, a hardware company EMC claimed world data to be 99 exabyte and predicted its growth rate to be 50% every year.
- Up to 500 videos are getting uploaded on YouTube each minute as per the statistic computed till May 2019.
- More than a trillion photos were captured in 2018, and that number is expected to rise by 7% in current year.
- In 2018, worldwide mobile traffic added up to 19.01 exabytes every month. In 2022, this data traffic is likely to reach nearly 77.5 exabytes every month globally, at a consolidated growth rate of about 46% annually. The measurements that used to be in gigabytes are now measured in zetta or even yottabyte due to the exponential growth of data.
- According to quite recent figures, Facebook generates approximately 4 petabyte of data every day. This biggest social networking platform has around 2.41 billion active users each month, according to the facts retrieved by quarter in 2019. It also has 400 users signing up for its service every minute; in that one minute, not only are accounts created, but also, 293 thousand statuses are created, 510 thousand comments are made, 136 thousand pictures are uploaded, and around 4 million posts get likes.

It's still ambiguous about how much data is generated every year, but the amount of information processed is no doubt huge. This immense volume of data needs different and distinct processing technologies, rather than relying on traditional processing and storage capabilities. At the end of the day, big data is too massive to be processed on any ordinary laptop processor; thus, we cannot analyse and store data using conventional database technologies. Now, distributed systems are utilized, and they store chunks of data at different places and can be brought back into single units by software. Collection and analysation of such vast amounts of data is a big challenge.

4.1.2 Variety – Different Forms of Data

Variety can be stated as different forms of data that can be used. Today's data is certainly different from data in past years, which was mostly structured data; such data includes bank statement information, including amount, date, account title, contact information, etc. Traditional data fits perfectly fine in any relational database. Big data handles diverse types of data, which can be divided into three main categories: structured, semi-structured, and unstructured data. Let's briefly discuss each one.

- *Structured Data* – Such data is well formatted and organized data, which has defined length for big data. Its examples are dates, numbers, combination of numbers, and words named as strings, etc. Such data is often stored in a relational database.

- *Semi-structured Data* – It lies between structure and unstructured data. It's a kind of structural data, which does not follow data model's formal structure linked to data tables or relational databases. Nonetheless, it contains markers or tags to detach semantic elements and impose records and fields hierarchies within data. Example include XML files and JSON; NoSQL also is considered a semi-structured type.
- *Unstructured Data* – Currently, most of the world's data (around 80%) reside in an unstructured category of data. It includes pictures, videos, updates on social media, such as tweets, statuses, posts, etc., voice recordings, CCTV footages; in addition to these, it also contains log files, machine data, click data, sensor data, etc. Unstructured data augment structured data, where things like audio files, web pages, MRI images, twitter feeds, and web logs are places. As evident, it contains everything that can be stored and captured, but it is not based on a meta model (collection of rules to surround an idea or concept) that precisely defines it. Unstructured data can better be defined when compared to structured data. Structured data can be thought of as data that is well defined under established rules; for instance, names are depicted as text, numbers will be used for money with minimum two decimal points, and there's a specific pattern followed by dates. Whereas in unstructured data, no rules are followed; for instance, a tweet, voice recording, or picture all are different but represent thoughts and ideas built on human understanding.

The variety of data types requires special algorithms with diverse processing capabilities. Organizing data to extract meaningful information is no ordinary task, particularly when its changing at a rapid pace. Big data technology's innovation and novelty have somehow enabled harvesting, utilization, and storage of all type of data.

4.1.3 Velocity – Speed of Data Generation

Velocity denotes the speed at which an enormous amount of data is getting generated, as well as created, collected, refreshed, and analysed. It's the incoming data frequency that needs processing. Twitter messages, SMS messages, swipes of credit cards, and Facebook status updates sent over specific telecom carriers at each minute, are generated at high velocity. One of the popular streaming applications that manages data velocity is a web service by Amazon called Kinesis. Few examples to apprehend the idea behind velocity are listed here.

- Google processes more than 70 thousand queries every second, making up to 4 million searches going on per minute, 240 million in each hour, and more than 5.76 billion searches each day. It reflects the phenomenon of change in our lives due to the internet.
- Facebook claims the per day incoming data rate to be more than 500 terabytes, for which it has a dedicated data warehouse at Prineville, Oregon. Some 240 billion pictures or more are stored by Facebook, in addition to

350 million new pictures being uploaded each day by users. To save these photos, storage of 7 petabyte gear each month is deployed by the data centre team of Facebook. Though it looks remarkable that Facebook stores data of more than 300 petabytes, a significant factor that should be accounted is the pace of creation of new data, aka its velocity. The speed with which data is increasing imposes the need for data analysis, but it also demands the data access and transmission rate to be prompt to enable real-time access to instant messaging, verification of credit cards, and access to websites. It's often argued that quick flow of information, as real time as possible, is a basic requirement of companies and that velocity can turn out to be more critical and important than volume since it provides greater competitive benefit. Its said that it's better to have limited data access in real time than low access rate to huge data. Availability of data at the right time helps suitable decision making in businesses; after a certain time period, that data may not be as significant as it was before.

4.1.4 Value – Data's Worth

By value, it means the worth of extracted data. An endless amount of data can turn into useless stock unless it has some value. The value characteristics of big data sit at the topmost position in a pyramid of big data, which signifies the ability to change data's tsunami into business.

The trade-off between cost and benefit for analysing and collecting data helps understand if it's monetary to reap the data, and it's the most crucial phase of embarking on true initiative of big data. Considerable value can be created in big data, which enables customers to understand well, targeting them according to that understanding, improving business performance, and optimizing methods. It's important to understand any strategy before embarking on it, along with its potential and challenges. The question arises whether gathered insights will help in creating a new line of product, cost-cutting measure, or cross-selling opportunity, or will it aid in discovery of crucial underlying effects that may produce a cure to a certain disease? If a company exploits big data correctly, after substantial investment on resources and time, then it can possess the ability to understand its customers, and, along with that, can monetize enormous information. That can allow the company to give offers that meet their customer needs at right time.

4.1.5 Veracity – Data Uncertainty

Veracity is trustworthiness, reliability, or quality of data. It states how accurate the data is. Take for an example, Twitter posts with abbreviations, hashtags, typos, etc., and the accuracy and reliability of such content. Another related example is GPS data usage, when users visit urban areas, GPS will sense off course. Tall buildings will bounce back satellite signals; so, in such situations, another source, such as road data, would be required to be fused with location data to provide correct information. This feature of big data is considered unfortunate since the increase in

any of the rest of the V's causes a drop in veracity; this may be similar to volatility and validity of big data (defined below). Gleaning loads of data is useless if it's not accurate.

Data-veracity knowledge enables better understanding of analysis, and risks associated with it, which eventually assists better decision making in business. Veracity ensures the accuracy and cleanliness of data, which keeps your system away from bad data accumulation by using various processes. Data with high veracity contains several records that are valuable for analysis and can contribute to complete results in a meaningful way, whereas data with low veracity contains a high proportion of meaningless data, often called noise. Medical trial or experiment data sets have high veracity.

4.1.6 Variability – Data Inconsistency

Variability in big data refers to few different concepts. First, it often means data inconsistencies, which are required to be found by outlier or anomaly detection techniques to do useful analytics. Secondly, variability in big data is because of dimensional data multitude, which is the result of numerous disparate data sources and types. Finally, it may state inconsistency in speed of loading big data to traditional databases. Variability is not the same as variety; say, for an example, four different coffee blends are offered by coffee shop, but if a customer gets the same blend each day, but finds it tastes different every day, then it's variability.

4.1.7 Visualization – Data Representation

Another important feature of big data is visualization challenge. Visualization discusses how data is represented through graphs or charts, etc. Representing huge quantities of complex data using charts and graphs, etc., is comparatively more effective than reports and spreadsheets packed with formulas and numbers to convey meaningful information. Existing visualization tools of big data are facing some technical difficulties because of in-memory technology limitations, poor functionality, scalability, and response time. To plot billions of data points, traditional graphs are not so useful; therefore, a need to represent data arises using various methods of data representation, such as tree maps, data clustering, parallel coordinates, cone trees, circular network diagrams, etc. Integrating this approach with a variable's multitude occurs from big data's velocity, variety, and complicated relationships among them.

4.1.8 Volatility – How Long to Store Data

How long can your data be kept until it is considered historic, irrelevant, or useless? How old should your data be to be considered useful? Before the arrival of big data, companies used to keep data indefinitely – a small number of terabytes may not produce excessive storage expenses; it may also be possible to store it in live databases without inducing performance issues. Due to volume and velocity of big data, volatility should be carefully considered as well. Volatility states how long the

data should be stored and still be valid. Real-time analysis on data requires the determination of the point at which data is not pertinent to ongoing analysis. A need to establish new rules for data availability and currency, as well as rapid information retrieval, emerges when the situation asks for it. These rules should be tied to business processes and needs with emphasis on complexity and cost of retrieval and storage process of big data.

4.1.9 Validity – Data Use

Like veracity, it also emphasizes the accuracy and correction of data for its expected use. As per Forbes, data scientists spend approximately 60% of their time on data cleaning before they even start analysis. Big data analytics are as essential as the underlying data; hence, it is desirable to adopt decent governance approaches to guarantee consistent quality, metadata, and definitions of data.

There are several other V's proposed by authors [16,17] such as viscosity, virality, vulnerability, vocabulary, and so forth. The above nine characteristics are further grouped into a few categories by authors. Veracity and variety come under data collection; velocity and volume come under the category of data processing; validity, variability, and volatility come under the category of data integrity. Alternatively, visualization goes in data visualization, and value is observed in data worth. These categories help in determining the gist behind each feature of big data.

4.2 DATA SOURCES OF BIG DATA

Big data consists of various types of cumulative data. Following are the key sources for big data: public, private, and community data; data wear out and self-quantified. Public data is the data seized by government institutions and local communities and transferred by business and management applications, such as transportation, energy use, and health care that required individual privacy under specific conditions. Private data are data contained by private companies, non-profit firms, and personal information that cannot be revealed by public sources; examples include consumer transactions, identification tags used by institutional supply chains, internet browsing, and usage of mobile phone. Data wear out is the type where data is collected from limited or zero-value data collection partners and attached with other resources for the purpose of creating new data. Another cause of such type of data is because of our information-seeking requirements for behaviours that can be utilized to conclude people needs or objectives. Community data is an unstructured data that includes consumer reviews on goods and voting review, for instance, twitter feeds, Facebook comments, and so on. Self-quantification is exposed from action and behaviours; for example, a digital watch customized to display exercise and movement. Such a self-quantification type of data helps to create a bridge between psychology and behaviours [18].

4.3 ARCHITECTURE

Big data design is the all-encompassing framework used to ingest and transform huge amounts of data with the goal to analyse the data for business decisions. The architecture is a blueprint for a solution of big data, grounded on the organization's business needs. Underneath headings, the layers involved in the architecture are precisely discussed.

4.3.1 Data Source

The organizations produce a huge proportion of data nearly once a day, and it is developing exponentially. The data-source layer integrates the data coming in from various sources, at various speeds, and in different associations.

4.3.2 Ingestion

Big data ingestion includes interfacing with different information sources, extracting the information, and distinguishing the transformed information. Following are the parameters for data ingestions: data velocity, data size, data frequency, and data format [19]. The data has been passed through from the following layers.

1. *Identification* – Arranged into different observed data structure; the unstructured data is entrusted with default structures.
2. *Filtration* – The data applicable for the end is constantly separated based on the (MDM) repository. MDM stands for Management.
3. *Validation* – After filtration, data is analysed in contrast of MDM metadata.
4. *Reduction of Noise* – By removing the noise and minimizing the inconsistencies, data is cleaned.
5. *Transformation* – Data is separated or combined based on its types, substance, and the necessities of the organization.
6. *Compression* – The size of the data is decreased without impacting its significance for the required method. It should be seen that weight does not impact the result assessment.
7. *Integration* – The refined dataset is united with the software utility layer, e.g, Hadoop.

4.3.3 Storage Layer

This is the key part for any big-data based system. It impacts the versatility, data structures and programming, and computational models of the system [20].

4.3.4 Staging

A staging is a middle of storage area utilized for data handling during the extraction, transformation, and loading (ETL) process.

4.3.5 Data Pipeline

The veracity of big data requires a quality of data coming in and out of the big-data processing pipeline. A big-data pipeline is divided into two parts: micro and macro pipeline. Micro pipeline work is based on level steps to make sub forms. A micro pipeline includes a granular data-handling step. Macro pipelines operate on workflow level and control each level workflow properly.

4.3.6 Data and Workflow Management

Big data architecture is built on large-scale distributed groups of data with scalable capacity. If a big data environment is built on a cloud, there is a need to spend time to establish a strong agreement with a cloud provider.

4.3.7 Data Access

The data structure exceptionally relies upon how applications or clients need to retrieve the information. Data-retrieval patterns should be known because some types of information can be redundantly recovered by enormous numbers of clients or applications (Figure 4.3). The big data architecture is shown in Figure 4.3 for the reader better understanding.

4.4 BIG DATA CHALLENGES

Big data is more like a torrential slide that's rapidly advancing down the mountain, getting faster and greater along the way, with a majority of organizations scrambling to keep pace with it. This is just like a skier requiring the fundamental equipment, such as helmet, gloves, etc., to survive an avalanche. Likewise, organizations need to be prepared for the avalanche, which, in this case, is big data; they need to be aware of essential tools while the avalanche is gaining steam. Having comparatively more opportunities to collect data from far more data sources is what makes data 'big'. Think about every one of the billions of gadgets that are currently internet-abled, cell phones and IoTs [21] sensors being just two cases. Though big

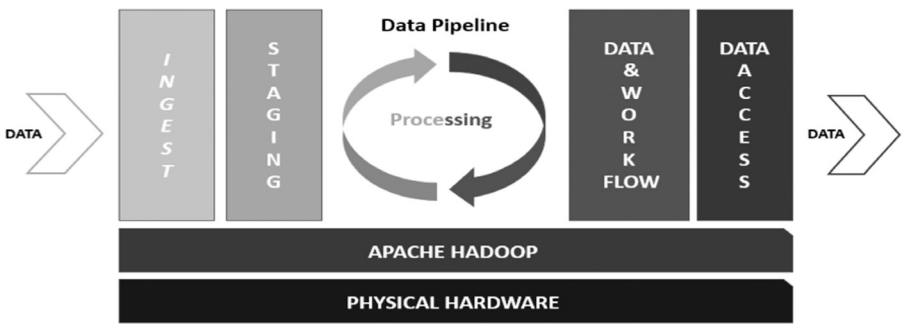

FIGURE 4.3 Big data architecture.

data offers several attractive opportunities, still professionals and researchers face many challenges while exploring large data sets and when they extract knowledge and value from such information. The complexity of these huge data sets poses plentiful difficulties at different levels, such as data storage, data capturing, sharing, searching, management, analysis, and visualization of data. Now imagine about all security-related problems that can arise. The idea of big data appeared from an unbelievable growth in the figure of IP-equipped devices. On a lighter note, big data is simply a term for entirely accessible data or information in a specified region that a corporation gathers with the objective of discovering trends or unseen patterns within it [22]. These, after getting discovered by analytical tools, can be utilized to yield an improved result not far off (more revenue, higher satisfaction level of customers, fast delivery of services, and so on). The other side of the coin states that the architecture for big-data storage reflects a new target of its security concerns for malware and illegal activities. The question is, should something happen to such an integral business asset, the outcomes could be disastrous for the association that accumulated it. Sadly, several big-data tools are open source, which are not often designed to keep security as a major function, prompting yet increased security issues. Often, the deluge of distributed streams and information surpasses our ability to harness.

Prior to going to battle, every general needs to study his enemies: the size of their army, type of weapons, battle count with results, their tactics, etc., in order to craft right strategy for battle and be prepared for the fight. Similarly, every decision-maker needs to comprehend what they are managing. Here, some major challenges attributed to big data are discussed below (Figure 4.4). Big data has a number of challenges and opportunities, as shown below.

4.5 BIG DATA ANALYTICS CHALLENGES IN BIG DATA

Big data carries huge transformative potentials and opportunities for different areas; but, at the same time, it additionally exhibits extraordinary challenges to binding such huge expanding volumes of data. Cutting-edge investigation is needed to interpret the associations among the features and examined data. For instance, data analysis empowers an association to fetch useful insights and keep track of the patterns that might influence business negatively or positively. There are various additional data-driven applications that require real-time analysis, such as social networks like Facebook, Twitter, LinkedIn, etc., area of finance, navigation, intelligent transportation systems like autonomous cars, astronomy, medicine, and so forth. Therefore, effective data-mining approaches, along with advanced algorithms, are required for accurate outcomes to observe the changes in different fields

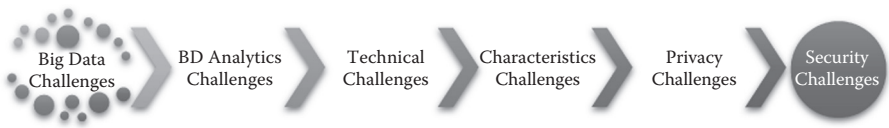

FIGURE 4.4 Major big data challenges.

and predict future perceptions [23]. Nonetheless, big data analysis is yet very challenging for numerous reasons: intricate nature of it, such as the Seven V's, the necessity for performance and scalability to scrutinize such enormous heterogenous data sets with realistic responsiveness [24–26]. These days, there are several analytical techniques including statistical analysis, data mining, visualization, and machine learning. Numerous studies deal with this area either by enhancing already proposed techniques or by proposing new or a tested ensemble of approaches or algorithms. Hence, big data accelerated the development of hardware, software, and system architectures. However, despite everything, there's a need of analytical progression to confront challenges of big data. Another issue is the assurance of timely response when the data volume is huge. In spite of the fact that scientists are trying to improve quality of data, as well as make analytical algorithms further robust (resistant to problems related to data), big-data analytics is not flawless or ideal. It's simply not yet possible to tackle some of the issues associated with data's reliability. Poor analysis leads to erroneous conclusions and correlations [23]. In the below section, major challenges encountered in big-data analytics are explored.

a. *Data Management Landscape's Uncertainty* – Due to the fact that big data is expanding day in and day out, new technologies and companies are being established every day. Finding out the best technology without initiating new problems or risk is a great challenge for companies [27].
b. *Big Data Capacity Gap* – Although big-data analytics is a rising field, the number of experts of this field is very limited. The reason behind insufficient experts is because it's a complex field and very few people understand the intricate and complex nature of this area. Hence, the talent gap presence in this industry is another big challenge for big data.
c. *Data attainment into big data platforms* – Companies are facing the challenge of handing this huge amount of rapidly increasing data. The variety and scale of today's data can overwhelm any data expert, which is what make it more crucial to make data easily available for managers.
d. *Synchronization among data sources* – Diversity of data sets increase the need to consolidate them into some analytical platform. In case this is overlooked, it can generate gaps, leading to incorrect insights and message [28].
e. *Important insights generation through big data analytics* – It is significant that organizations obtain legitimate insights from big-data analytics, and it's also equally important to right department approaches to this data and access it. Thus, it's a major challenge to bridge this gap in productive fashion.
f. *Biased markers in Big Data Analytics* – The big-data algorithms are frequently grounded on explicit markers associated with the analysed element. And because of this, analytics can be misrepresented. When someone cracks what markers impact the results, they can modify the analysed element to satisfy the prerequisites set by the markers. As stated, that big-data analytics depend on specific markers of the studied item. If the individual appending the marker is biased toward the matter, it will influence the outcome. Hence, biased markers result in biased analysis.

4.6 TECHNICAL CHALLENGES IN BIG DATA

It tends to be anything but difficult to get lost in the assortment of big data technologies currently available in the market. Do you require Apache Spark, or would the pace of MapReduce be sufficient? Is it desirable to use HBase or Cassandra for data storage? Discovering answers to these questions can be complicated. And it's easy to pick inadequately, on the off chance that you are searching the ocean of technological prospects without a clarity of what you require [23].

These technical difficulties are further divided to subheadings defined below.

a. *Fault Tolerance* – The arrival of new technologies such as big data and cloud computing increase the expectations that every time a failure occurs, the harm done should be within tolerable threshold as opposed to starting the entire task from scratch. However, fault tolerance is incredibly hard computing, involving complex algorithms. It's basically impractical to devise completely foolproof and 100% reliable machine or software. Thus, the fundamental assignment is to decrease the likelihood of failure/error to an acceptable stage. It is unfortunate that the more we struggle to decrease this probability, the higher the expense. There's a strategy to overcome this problem. It divides the entire calculation being performed into smaller tasks and allocates these tasks to various nodes for the sake of computation. One of the nodes is given the responsibility to check proper working of these nodes. In case something happens, that specific task is restarted. However, there can be a scenario where it's not possible to divide the entire computation into independent tasks. The tasks can be recursive where the input of previous task depends on the computation of next one. So, it can be cumbersome to restart the whole computational task [23].
b. *Scalability* – The technology behind the processor has changed recently. The speed of clock has slowed down to a great extent, whereas the number of cores has increased in a processor. Earlier data-processing frameworks needed to stress over parallelism across nodes in a specific cluster, but recently the concern has moved to intra-node parallelism. The techniques used in past for parallel-processing across nodes aren't efficient enough to handle parallelism within a single node. It's because considerably more hardware resources are jointly used across a core in a separate node. The issue of scalability in big data has led to cloud/distributed computing, which presently aggregates many dissimilar workloads with fluctuating performance objectives into huge clusters. This necessitates a higher level of resource sharing, which is costly and furthermore carries with it many challenges, such as how to execute and run different jobs so we can cost-effectively achieve the goal of every workload. It additionally requires managing the failures in an efficient way that arises more often than if operating on big clusters [29].
c. *Data Quality* – Huge amount of data collection and storage emanates a substantial cost. Better results can be attained if more data is fed to predictive analysis and decision-making processes in any business. Business leaders always want more data storage; on the other hand, IT leaders consider

technical aspects before data storage. Big data essentially focuses on quality of data storage, rather than having irrelevant bulk data, in order to draw better outcomes and conclusions. This further prompts different queries, like how it very well maybe guaranteed what data is relevant, how much of the data would be sufficient for the process of decision-making, and even if the data stored is correct or how not to reach inference from it, etc.

d. *Data Heterogeneity* – Unstructured data embodies nearly every sort of data being constructed, for instance, recorded gatherings, PDF documents handling, social media interactions, emails, fax transfers, and many more. Structured data is constantly being organized in extremely automated and manageable ways. It depicts proper integrations with a database; however, unstructured data is totally crude and disorderly, So, working with such data is obviously exorbitant and cumbersome. Conversion of such unstructured data to organized data is not feasible, either. Digging through this data is unwieldy as compared to structured data that be easily managed [24].

4.7 CHARACTERISTICS-ORIENTED CHALLENGES OF BIG DATA

Every property of big data possesses a challenge, few of them of major concern are discussed below.

4.7.1 Data Volume

The very first issue associated with this property is related to storage. The amount of space increases with the increase in data volume, so there is a need to store it efficiently. Apart from the fact that gigantic volumes of data should be retrieved at a quick pace to fetch output from them, there are other areas that need attention, as well such as storage costs like cloud storage versus in-house storage of data, bandwidth, networking etc. [30]. With the expansion in data volume, the worth of data records reflects decreases in proportion to richness, age, type and quality [31]. The advent of social-networking websites has led to a generation of data of the order of peta/terabytes each day. Aforementioned data volumes are hard to tackle using existing conventional databases [31].

4.7.2 Data Velocity

More and more data is being generated on a regular basis by computer systems, and both analytical and operational speeds as well as the amount of consumers of said data are growing. Individuals need the entirety of data, and they need it at the earliest possible time, prompting to what is popular as high-velocity data. Data of high velocity denotes millions of columns of data every second. Classic outdated database systems are not fit enough for executing analytics on such data, which is continuously in motion. Unfortunately, the state-of-art technology can't manage data that restrict the collection of data, such as data generated by machines and human activities like Twitter feeds, log files, website clicks, and mortar, etc. [31].

4.7.3 Data Variety

Big data comes in numerous fashions, namely messages, images, and updates in social-networking sites. Global-positioning systems signals come from smartphones, sensors, and many more. A considerable number of these big data sources are almost new or somewhat as old as social-networking websites similar to Facebook and Twitter launched in 2004 and 2006, respectively. Cellular phones and other such devices can be put in the same class. The ubiquitous nature of these devices makes traditional databases unsuitable for these data. Wide range of these data is unwieldy, noisy, and unstructured, which entails thorough techniques for data-based decision-making. Hence, improved algorithms for analysing such data is a critical issue [32].

4.7.4 Data Value

Companies store data to gain useful insights from it and utilize it for business-intelligence analytics. This process of storing creates a gap between IT professionals and business leaders. As mentioned before, business experts are found concerned over data value for their business and how can the data be profitable for them. They believe that more data generates more insights, whereas IT experts also deal with technical details when storing and processing big data [31].

4.8 PRIVACY CHALLENGES

Although people appreciate the ease brought by big data, they experience numerous inconveniences as well. If big data is not very safe for user data during the time of use, it will legitimately undermine the protection of users and data security. As per many protection contents, it tends to be subdivided into categories such as anonymous protection, anonymous identifiers, and privacy protection. A person's privacy is a very important and sensitive issue that holds technical, conceptual, and legal significance [33]. The private data of a person when joined with external huge datasets leads to new facts inference about that individual; for example, vendors posting relevant ads after observing users' spending habits, such as on specific designer clothes, styles, or locations, etc. It is possible that the individual may not want the data owner or any other person to know these secret facts. Currently, numerous organizations believe that the identifiers will be covered up, after processing the information anonymously. However, security assurance can't be viably accomplished through the use of anonymous protection. Users information is gathered and used so as to increase value to organization's business. It's made possible by forming insights in users' private lives, about which they are not aware. Another significant consequence emerging is called social stratification where an aware and literate person takes benefits from big data's analytical ability [34] and the underprivileged or naïve users will be identified easily, hence treated awfully. Many countries lack regulations and rules for management of user information in the current era and don't have decent supervision systems. All this coupled with user's unawareness of self-protection has instigated many issues because of leakage of information [35].

4.9 SECURITY CHALLENGES

Security and privacy are huge challenges in the era of big data, and this concern is increasing with the growth of data each second [36]. A key reason behind this is that information is now easily and widely accessible. Professionals from various areas, such as medicine, business, and government, are sharing data on a large-scale with each other. Yet, the developed tools and technologies in use until today are not competent enough to handle such enormous amount of data and are not efficient enough to give satisfactory security [37]. The technologies are deficient to provide enough maintenance features for privacy and security due to the absence of a fundamental understanding about ways of providing security to such an enormous bulk of data and because necessary training is not offered concerning how to equip these large-scale data sets with security and privacy [37]. Big data security, together with privacy maintenance in regard to big data, lacks suitable policies that guarantee compliance with recent approaches in security and privacy. Current technologies are continuously encountering accidental and intentional breaches because of having weak privacy and security-maintenance abilities. Therefore, it's necessary to update and reassess present methods to prevent continuous data leakages. It has also been realized that IT dompanies spend a minuscule financial resource on big data protection. Nearly 10% of an organization's IT-related budget ought to be spent on security, but less than 9% on average is spent, consequently making it difficult for these companies to ensure protection of data [37].

4.10 CURRENT SECURITY CHALLENGES IN BIG DATA

Before digging deep into security concerns of big data, let's first define big data security to get better understanding of each aspect associated to this subject.

4.10.1 Big Data Security – A Definition

Big data security is a joint term for every tool and measure used to protect data, as well as analytics processes from theft, attacks, or other similar malicious practices that can in any way harm or adversely influence them. Similar to other types of cyber security, variants of big data are also concerned with offenses that get initiated either from offline or online domains. For organizations working on cloud, this challenge is multifaceted. These dangers include online information theft, DDoS attacks, server crashes, and ransomware. The matter can be much more problematic when organizations store confidential or sensitive information, such as credit card numbers, customer personal information, or even just contact particulars. Furthermore, these attacks on big data storage organizations could originate serious financial aftermaths like fines, losses, sanctions, or litigation expenses, etc.

Big data is nothing novel to bigger organizations. However, its popularity is not restricted to large organizations; instead, it equally holds the same status among small and medium-size firms. The reason behind it is its simplicity in terms of management; along with this, it comes with the benefit of reduced cost [38]. Cloud storage has encouraged the use of data collection and data mining. Be that

as it may, this big data integration with cloud-based storage has instigated challenges to security and privacy threats. The cause behind such breaches is probably because that the applications are designed to accumulate a certain quantity of data, but they are unable to define a mechanism of handling such a gigantic volume of data that the previously mentioned data sets have. Similarly, these security technologies are not efficient enough to cope with dynamic data and are only suitable to control the static data. While the common public reaps the benefits of big data, they also confront numerous inconveniences. If the data of users is not protected well, then it results in security and privacy concerns. A few security case studies are discussed in the upcoming section, and after that, challenges are discussed.

4.10.2 Case Studies of Security Breaches Depicting Their Impact on Organizations

According to Cyber Edge [39], more than 60% of 763 security professionals participating in a study disclosed effective cyber-attacks on their medium to large corporations. A report in 2015 by Data Breach Investigation corresponded about 80K security occurrences, in addition to 2.122K affirmed data breaches [40]. A couple of years ago, a solitary data breach affected around 1–10 million records of a target company but these days, even a single data breach can result in a compromise of around 0.2 billion records, triggering multimillion-dollar loss and harm to brand names, alongside confronting governing penalties. Consider Ashley Madison [41], a site for extramarital affairs, which endured a data breach in 2015 and suffered data leakages of 25 gigabytes. This not only caused harm to the reputation of this social website but also brought about disturbances in clients' lives, with reports of two suicide cases. One more instance is of Target [42], a discount retailer in the United States, through a data rift occurring during Christmas season in 2013. The breach occurred at POS (point of sale) system where hackers stole information of about 0.04 billion debit and credit cards. In addition to that, they also gained unauthorized to data of around 70 million consumers, which included user names, email addresses, phone numbers, etc. A report by Naked Security stated that this data break had cost about $290 million, of which insurance charges covered $90 million. In spite of that, the company is still facing lawsuits by the shareholders and inquiries by state attorney and the Federal Trade Commission, which could further cost them over $30 million, collectively a loss of $300 million.

Take another example of New Jersey based Horizon Blue Cross Shield (HSC) [43]. The company encountered two similar cases, in 2013 and 2016, respectively. In January 2008, the company modified its corporate policy because of the theft of an employee's laptop. The policy stated to implement encryption on all devices such as mobiles, desktop systems, and laptops. Even then, in 2013's act of data breach, many laptops were found unencrypted by the investigating officials. Again, another data leakage occurred and the company realized that laptops were protected by passwords, but had no encryption implemented on them. This breach exposed customer names, birthdates, SSNs, insurance identification numbers, and clinical

information. This made the company to pay $1.1 million for inferior security solutions. Comparable case is of Yahoo [44–46]. A data breach happened around 2013–2014, which was reported in September 2016 when Yahoo was in negotiations with Verizon. They reported that they suffered the biggest breach ever in the history, where about 3 billion user accounts were sacrificed. Around the same time, 117 million users lost information on LinkedIn [44,47], a networking website of professionals, when it was hacked and a massive number of email addresses and passwords were purchased by the black market. Another biggest data breach is of Facebook [48], where nearly 50 million user profiles were collected for Cambridge Analytica with intention to influence voters' choice in the US election. The the analytics firm collaborated with current US President Donald Trump's team. The Cambridge Analytica team, along with Steve Bannon (personal advisor) used private information of users in an unauthorized manner to construct a system that would profile each voter to target them with custom political ads. In a very recent case in March 2019, Toyota issued an official statement on their website confirming a data breach, which hypothetically affected around 3.1 million users. Several other cases of data breaches that affected large population of users have occurred; only a few of them are stated above, but even from these examples, it is evident that big data can cause serious security problems. It is by these bitter experiences that IT and business specialists are learning to improve security and protect big data.

4.11 MAJOR SECURITY ISSUES OF BIG DATA

Here's a list of few of the security challenges [49] that may occur while using big data.

4.11.1 Distributed Frameworks Security

The majority of big data applications distribute massive processing jobs over several systems for rapid analysis. A eminent example of it is Hadoop, which is an open-source technology and initially had no security of any kind. Distributed and shared processing may point to the idea that less data would be processed by any of the system, yet it implies significantly more systems where security problems can manifest.

Computational security and other advanced resources in a distributed environment like MapReduce [50], a processing technique of Hadoop, tend to be deficient in security protection. MapReduce framework splits an input file into numerous lumps, and afterward, a mapper for each piece interprets the data, performs computations, and produces outputs as key-value pairs. Another actor, named reducer at that point, merges the values attached with every unique key and yields the outcomes. The chief fears here are: verification of the mappers and protection of data from a malevolent mapper. Mappers sending inaccurate outcomes are too tough to spot and inevitably bring about erroneous aggregate results. With very large data sets, malicious mappers are too hard to be detected as well, and they eventually damage essential data. Leakage of private records by mappers, be it intentional or accidental, is also an alarming matter. MapReduce calculations are frequently

endangered by attack, such as man-in-the-middle, replay, and denial-of-service attack. Rogue nodes can be made part of the cluster and in order obtain replicated information or send modified MapReduce code. Making snapshots of authentic nodes and introducing their duplicate copies again is a simple attack in virtual environment and clouds, whereas it's difficult to detect it [32]. The two fundamental prevention approaches for it are to securely verify the mappers along with security of information within the sight of unauthorized and unlawful mapper.

4.11.2 Nonrelational Data Stores Protection

Nonrelational databases, such as NOSQL databases adapted to store enormous volume of data, tackle several challenges associated with big data analytics without worrying a lot over security problems. NoSQL databases provide no clear security implementations, instead comprising security implanted in the middleware. Maintenance property transactional integrity is pretty negligent in these databases. Multifaceted integrity constrains can't be instilled in NoSQL datum as it hampers with its operational ability of giving improved scalability and performance. These databases have relatively weak authentication and password storing mechanisms. Its because they utilize basic or digest-related authentication (HTTP), and henceforth are vulnerable to the man-in-the middle attack. Also, representational state transfer is prone to injection attacks like JSON injection, array, view, REST, generalized query language, schema injection, and many more, including cross-site request forgery and cross-site scripting. NoSQL does not also support blocking with aid from third parties. Another limitation in terms of authorization approach is that it provides authorization only at higher layers. Consequently, it facilitates permission on each database level instead of providing it at data collection level. NoSQL databases are exposed to attack within it due to permissive security structures. Poor log analysis and logging methods leave these flaws unnoticed along with many other principal security mechanisms [32].

4.11.3 Storage Security

Architecture of big data store data on multiple layers, conditional to business requirements for cost versus performance. For example, hot data of high precedence will typically be stored on smart flash medium. Therefore, confining the storage will be equivalent to creating tier-cognisant strategy.

A multitiered storage medium was used to keep transactional and data logs. The increase in data size, the issue of scalability, and availability led to the use of auto-tiering for storage of big data. But auto-tiering comes with a limitation: it does not record location of data, unlike past approaches of multitiering storage where the IT professional would exactly know where the data was and when was it placed there. This introduced new challenges for security of data storage. Just to state one, the service providers often look for hints that assist them to correlate user data sets and their activities and become acquainted with certain properties that can prove to be decisive for them. However, it's not possible for them to disrupt into data by conquering the encipherment. As the cipher text is stored by data owner in an auto-storage

framework, he circulates the private key to every user and gives the right of data access to certain parts of the system to specific users. The unreliable service provider is not an authorized user of the key, but he might plot with other users by trading the key and data; subsequently, he can get access to data he is not legally allowed to access. If the environment is multitiered, service providers can promptly roll back an attack on set of users or deliver an obsolete form of data while the recent data is uploaded by then in the database. Data loss and data tempering came about by malicious users, regularly leading to conflicts among the users or between the providers [32].

4.11.4 Monitoring Real-Time Security

Real-time security checking has been a progressing challenge in the case of big data analytics, primarily because of the amount of alerts security devices generated. These signals may or may not be correlated, but they can cause numerous false positives; because individuals lack the ability to effectively manage such an enormous amount of them at such pace, they are ignored or clicked away [51]. Monitoring of security necessitates that infrastructure or big data platforms be intrinsically secure. Several threats are posed to the infrastructure, which include web application threats, rogue administrative access to nodes or applications, and intrusion on the line. The security of each component of the infrastructure and security after their integration must be examined. Let's take the example of Hadoop, the cluster execute in a public cloud, hence the cloud security, which itself is an ecosystem of quite a few components comprising of storage, computing, and network mechanisms; it needs to be carefully studied. The security of Hadoop nodes, their interconnections, data storage, and the cluster in its entirety should be weighed. Also, monitoring application security containing pertinent associations that ought to pursue secure coding standards must be investigated as well, and the input sources from where the information originates must be accounted.

4.11.5 Privacy-Preserving Data Analytics and Mining

Big data is vulnerable to privacy appropriation, decrease of civil freedom, obtrusive marketing, and an upsurge in corporate and state control. A worker in an organization responsible for storage of big data can possibly abuse power and disregard privacy guidelines. For instance: Big data workers can stalk individuals by observing conversations, if they work for a social networking-based company that aids chatting. An untrustworthy partner in business can invade personal data and transfer it into the cloud since policies permits the owner to handle cloud infrastructure [32].

4.11.6 Granular Audit

The real challenge is when the monitoring system receives notification at the very instant an attack happens. There exists a possibility of frequent new attacks or unexploited true positives. To find a missed attack, required information is termed as audit information. Audit information from any gadget must be finished, or it must present us facts concerning what precisely occurred and what turned out wrong. It

must deliver on-time access, with the goal that it obliges the need of compliance, forensic examination, rules, and regulations. It must not be meddled with and should be available only in permitted areas [32].

4.11.7 End-Point Security

Companies gather data from diverse sources, including software applications, hardware equipment, and end-point tools. Validation and the data collection from diverse sources is a serious challenge. Malicious users tamper with the target machine from where collection of data happened or temper the application responsible for collecting data, configuring the device to float malicious data as input to the core system of data collection. Counterfeit IDs are made by malignant users who intend to deliver malicious data as input to the data-collection system. Sybil attacks are a prominent ID cloning attack in a BYOD (bring your own device) setting where a vindictive user brings his own machine, forged as an authorized machine, and injects malicious input to the system. Sensory-data input sources can also be controlled, like artificially varying the temperature from its sensor and inserting malignant input into the collection process of temperature. Global positioning systems (GPS) can be altered similarly, where a user may modify the data during its transmission from source to the chief data-collection system. It can somehow be categorized under a man-in-the-middle attack [32].

4.11.8 Data-Centric Security Based on Cryptography

To control data visibility to individuals, systems, and organizations, two basic methodologies are commonly used. The first is to limit access to primary systems like hypervisor or operating systems. The next one is embodying the data in a defensive shield due to cryptography. The initial approach gives a bigger surface for attack. There exist numerous attacks, such as buffer overflow and privilege-based escalation attacks that circumvent access-control applications and gain data access. Shielding data from end to end using encryption support considerably reduces the attacking surface. Although it seems like an impossible task, it is susceptible to translate side attacks and fetch secret keys. Different threats related to cryptography-based, access-control strategy-exploiting encryption are: it ought not be recognizable by the enemy, and the equivalent plain-text data observes the cipher text, regardless of whether he needs to pick between right and wrong plain text. For a searching and refining of encrypted data, the cryptography protocol may not allow the adversary to learn anything regarding encrypted data past the relating predicate, regardless of whether fulfilled or not. It must also guarantee that the attacker must not have the option to construct information that originated from asserted sources; this probably would be bogus influencing data integrity [32].

4.12 SOLUTIONS TO SECURITY CHALLENGES

Several protections against security challenges have been proposed [52]. Few of theoretical solutions are discussed below.

4.12.1 Complete Data Supervision of Social Networks

In the era of big data, creation of online media has revolutionized interpersonal communication. Establishing strong supervision of data is critical. First, it is mandatory to fortify the management and supervision of data and secretly ensure the protection of data at network for mysterious social media. Second, conduct management and supervision of social data to safeguard private information security, and make sure this information is not exploited by the cyber criminal. Besides, to enhance users' knowledge of precautionary measures for safety and to mitigate personal information filling, vigilance drawbacks and self-prevention understanding are also required. At last, government bodies should announce better policies and regulation for big data applications at the earliest possible opportunity.

4.12.2 Improvement in Legal Mechanism

With the expansion of society, the public pays more attention to privacy, and governing bodies give more consideration to the safety of discrete rights of each citizen and present many counter measures for information protection. In the criminal law amendment, the principles for the defence of citizens' private information are stated explicitly; that is, come what may, the public officer sees about the resident's data, he/she must not utilize any resources to deliver information to others. If the information is leaked, it must accept some legal responsibility. Therefore, to protect the big data security, the administration needs to start a comprehensive personal information protection law to safeguard national's individual data.

4.12.3 Improvement to People Awareness of Data Quality

With the constant development in the era of big data, the amount of data has amplified naturally. People must adjust to variations in the times and progressively improve their data literacy and awareness. Data literacy is primarily designed for research scientists and public servants. Data awareness is focused at the overall public and need inhabitants to comprehend the significance of big data. Do not randomly print information regarding your own confidentiality on the internet, and do not publish others' information to avoid any exploitation by criminals.

4.12.4 Put Security First

While designing solution architecture, put security at the top-most priority. Being careful and aware of possible security risk at every stage allows IT professional to resolve the issues concurrently.

4.13 CONCLUSION

Big data came out from this incredible escalation in the number of IP-equipped end points. It's a fundamental term for all available data in any area that a corporation can collect with the aim of finding secret trends and patterns. These with the help of

analytical tools can yield better results in the form of greater revenue and user satisfaction. Alternatively, the other side of the coin is that big data architecture, particularly its storage, depicts a new target of security issues for malware and criminal activities. Security and privacy are among the most talked about challenges of big data. Security breaches examples are there to showcase how data breaches can affect not only a firm's reputation but also its financial stability, along with sacrifice of users' confidential information. Proper policies are required to ensure security in big data, not only that users need to increase their awareness as well. Also, many data-mining and machine-learning algorithms, such as deep-learning models, can be exploited to reduce this problem.

REFERENCES

[1] Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 42–47). IEEE.
[2] Almadhoor, L. (2021). Social media and cybercrimes. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(10), 2972–2981.
[3] Humayun, M. (2020). Role of emerging IoT big data and cloud computing for real time application. *International Journal of Advanced Computer Science and Applications*, 11(4), 1–13.
[4] Bekker, A. (2017). *Big data: Examples, sources, and technologies explained*. [online] Scnsoft.com. Available at: https://www.scnsoft.com/blog/what-is-big-data [Accessed 29Jul.2019].
[5] Humayun, M. (2021). Industry 4.0 and cyber security issues and challenges. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(10), 2957–2971.
[6] Zikopoulos, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.
[7] Alshammari, H. (2019). U.S. Patent No. 10,268,716. Washington, DC: U.S. Patent and Trademark Office.
[8] Laney, D. (2001). 3-D data management: Controlling data volume, velocity and variety. *META Group Research Note, 6*, 70.
[9] Dijcks, J. (2012). Big data for the enterprise. Oracle report.
[10] Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). Analytics: The real-world use of big data. IBM report, 1–20.
[11] Katal, W., & Goudar. (2013). Big data: Issues, challenges, tools and good practices, contemporary computing (IC3). In Sixth 2013 IEEE International Conference on Noida. IEEE.
[12] Krishnan, K. (2013). *Data warehousing in the age of big data*. Elsevier.
[13] Khurshid, K., Khan, A., Siddique, H., & Rashid, I. (2018). Big data-9Vs, challenges and solutions. *Technical Journal*, 23(03), 28–34.
[14] Lomotey, R. K., & Deters R. (2014). Towards knowledge discovery in big data. In *2014 IEEE 8th International Symposium on Service Oriented System Engineering*. IEEE.
[15] Boyd, D., & Crawford, K. (2012). Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society,* 15(5), 662–679.
[16] Arockia Panimalar, S., Varnekha Shree, S., & Veneshia Kathrine, A. (2017). The 17 V's of big data. *International Research Journal of Engineering and Technology*, 4(9), 329–333.

[17] Shafer, T. (2019). The 42 V's of big data and data science. Available at: https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html.
[18] George, G., Haas, M. R., & Pentland, A. (4 April 2014). Big data and management. *Academy of Management Journal*, 57(2). 10.5465/amj.2014.4002.
[19] Gill, N. S. (2017, Mar 03). Data ingestion, processing and architecture layers for big data and IOT. Available at: https://www.xenonstack.com.
[20] Sakr, S., & Gaber, M. (2014). *Large scale and big data: Processing and management*. Auerbach Publications.
[21] Kaul, L., & Goudar, R. H. (2016, November). Internet of things and big data-challenges. In 2016 Online International Conference on Green Engineering and Technologies (IC-GET) (pp. 1–5). IEEE.
[22] Lancaster, T. (2019, January 25). 9 Key big data security issues. Available at: https://www.alienvault.com/blogs/security-essentials/9-key-big-data-security-issues.
[23] Zaman, N., Seliaman, M. E., Hassan, M. F., & Márquez, F. P. G. (2015). *Handbook of research on trends and future directions in big data and web intelligence*. Pennsylvania: Information Science Reference.
[24] Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, 30(4), 431–448.
[25] Nyunt, K., & Noor, Z. (2015). The effectiveness of big data in social networks. In *Handbook of research on trends and future directions in big data and web intelligence*, pp. 362–381. IGI Global.
[26] Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2016). Big data analytics. In *Big data technologies and applications,* pp. 13–52. Cham: Springer.
[27] Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, 70, 263–286.
[28] Ghosh, K., & Nath, A. (2016). Big data: Security issues and challenges. *International Journal of Research Studies in Computer Science and Engineering*, 1–9.
[29] Bertino, E., & Ferrari, E. (2018). Big data security and privacy. In *A comprehensive guide through the italian database research over the last 25 years,* pp. 425–439. Cham: Springer.
[30] Thayananthan, V., & Albeshri, A. (2015). Big data security issues based on quantum cryptography and privacy with authentication for mobile data center. *Procedia Computer Science*, 50, 149–156.
[31] Matturdi, B., Zhou, X., Li, S., & Lin, F. (2014). Big data security and privacy: A review. *China Communications*, 11(14), 135–145.
[32] Big Data Working Group. (2013). *Expanded top ten big data security and privacy challenges*. Cloud Security Alliance, 1–39.
[33] Abdrabo, M., Elmogy, M., Eltaweel, G., & Barakat, S. (2016). Enhancing big data value using knowledge discovery techniques. *IJ Information Technology and Computer Science*, 8, 1–12.
[34] Tene, O., & Polonetsky, J. (2011). Privacy in the age of big data: A time for big decisions. *Stanford Law Review Online*, 64, 63.
[35] Soria-Comas, J., & Domingo-Ferrer, J. (2016). Big data privacy: Challenges to privacy principles and models. *Data Science and Engineering*, 1(1), 21–28.
[36] Suganthi, M. (2018). Big data: Security issues, challenges and future scope. *International Journal for Research in Science Engineering & Technology*, 5(1), 10–20.
[37] Schmitt, C., & Shoffner, M. (2013). Security and privacy in the era of big data. *The SMW, a Technological Solution to the Challenge of Data Leakage*, 1(2).

[38] Kshetri, N. (2014). Big data's impact on privacy, security and consumer welfare. *Telecommunications Policy*, 38(11), 1134–1145.
[39] Cyber Edge Group. (2014). *2014 Cyberthreat Defense Report*. Cyber Edge Group.
[40] Alayda, S. (2021). Terrorism on dark web. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(10), 3000–3005.
[41] Mansfield-Devine, S. (2015). The Ashley Madison affair. *Network Security*, 2015(9), 8–16.
[42] Humayun, M., Niazi, M., Jhanjhi, N. Z., Alshayeb, M., & Mahmood, S. (2020). Cyber security threats and vulnerabilities: A systematic mapping study. *Arabian Journal for Science and Engineering,* 45(4), 3171–3189.
[43] Date, F. C., & Date, R. (2016). *Horizon Blue Cross Blue Shield of New Jersey*. https://www.horizonblue.com/sites/default/files/2017-05/horizon-bcbsnj-2016_annual_report.pdf.
[44] Ullah, A., Azeem, M., Ashraf, H., Alaboudi, A. A., Humayun, M., & Jhanjhi, N. Z. (2021). Secure healthcare data aggregation and transmission in IoT—A survey. *IEEE Access,* 9, 16849–16865.
[45] Thielman, S. (2016). Yahoo hack: 1bn accounts compromised by biggest data breach in history. *The Guardian*, 15, 2016.
[46] Trautman, L. J., & Ormerod, P. C. (2016). Corporate directors' and officers' cybersecurity standard of care: The Yahoo data breach. *American University International Law Review*, 66, 1231.
[47] Layton, R., & Watters, P. A. (2014). A methodology for estimating the tangible cost of data breaches. *Journal of Information Security and Applications*, 19(6), 321–330.
[48] Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*, 17, 22.
[49] Terzi, D. S., Terzi, R., & Sagiroglu, S. (2015, December). A survey on security and privacy issues in big data. In 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST) (pp. 202–207). IEEE.
[50] Almrezeq, N. (2021). "An enhanced approach to improve the security and performance for deduplication." *Turkish Journal of Computer and Mathematics Education (TURCOMAT),* 12(6), 2866–2882.
[51] Singh, R., & Ali, K. A. (2016). Challenges and security issues in big data analysis. *International Journal of Innovative Research in Science, Engineering and Technology*, 5(1), 257–264.
[52] Zhang, D. (2018, October). Big data security and privacy protection. In *8th International Conference on Management and Computer Science (ICMCS 2018)*. Atlantis Press.